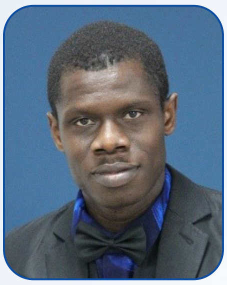

Effective Medical Data Curation for the Optimum Performance of Machine Learning and Deep Learning Models

Chukwuebuka Joseph Ejiyi[1], Zhen Qin[1], Makuachukwu Bennedith Ejiyi[2]

[1]School of Information and Sofware Engineering, University of Electronic Science and Technology of China

[2]Faculty of Pharmaceutical Sciences University of Nigeria Nsukka

Abstract

Although oftentimes not considered very crucial, effective preparation of data for implementation in Machine Learning and Deep Learning remains the fulcrum upon which the performance of the models is hinged. The procurement of data is the very first important step in every Artificial Intelligence project but in the medical field and in many other fields, the procurement of data is usually difficult. This is in some cases a result of some privacy and other related policies. Although with the availability of technology and smart devices lots of data is generated which tends to arrest the challenge of data procurement. With this machine-generated data, comes anomalies that may be attributed to the devices themselves or individuals, additionally, when the data is manually generated, the possibility of missing data and more is usually high. In the medical field, some patients may not be willing to give some information or may give the wrong ones thereby prompting the need for the curation of data. Several steps before curation are required which are not limited to the formulation of the problem and the proper data collection itself but may include data exploration and many more. In the task of curation, the data is basically cleaned and validated using various tools which are geared towards making the data understandable by machines and also make for easier implementation for both Machine learning. For the purpose of predicting cardiovascular diseases for which we used Shapely and other data curation methods including data augmentation, the following improvements were recorded from the data 11.25%, 11.64%, 10%, and 10.46% respectively for accuracy, sensitivity, F1 score and precision using gradient boosting algorithm while the ANN gave 100% accuracy on a UCI heart disease dataset. This is an indication that the use of data curation goes a long way in improving the performance of both machine learning and deep learning models.

Biography

Chukwuebuka Joseph Ejiyi received his Bachelor's Degree in 2014 from the Federal University of Technology Owerri (FUTO) Nigeria. He went on to obtain a master's degree in Software Engineering at the University of Electronic Science and Technology of China (UESTC) in 2021 where he majored in deep neural networks. He is currently pursuing a Ph.D. degree with the School of Information and Software Engineering at UESTC Chengdu China. His research interest is in Artificial intelligence, Deep Learning and he is interested in Object detection using a single-stage neural network as well as image classification/segmentation and is currently working on image/data analysis, especially with regard to the medical field.